Data Mining - Homework 1

Howie Benefiel *phb337*

February 18, 2019

Problem 1.

(a) The probability that $X = 1$ is ${}^{1}\!/_{4} + {}^{1}\!/_{3} = {}^{7}\!/_{12}$.

(b) The probability that $X = 1$ conditioned on $Y = 1$ is $\frac{1/3}{1/2} = {}^{2}\!/_{3}$.

(c) The variance of X is $E[X^2] - E[X]^2$. Breaking it down:

$$E[X] = (1/4 + 1/6) \cdot 0 + (1/4 + 1/3) \cdot 1 = 7/12$$
$$E[X^2] = (1/4 + 1/6) \cdot 0^2 + (1/4 + 1/3) \cdot 1^2 = 7/12$$

So $Var(X) = 7/12 - 7/12^2 = 49/144$.

(d)

The variance of X conditioned on $Y = 1$ is $E[X^2|Y = 1] - E[X|Y = 1]^2$.

$$E[X|Y = 1] = \sum_{x=0}^{1} x \cdot \frac{P(X = x\, Y = 1)}{P(Y = 1)}$$
$$E[X|Y = 1] = 0 \cdot \frac{1/4}{1/2} + 1 \cdot \frac{1/4}{1/2} = 1/8$$

and

$$E[X^2|Y = 1] = \sum_{x=0}^{1} x^2 \cdot \frac{P(X = x\, Y = 1)}{P(Y = 1)}$$
$$E[X^2|Y = 1] = 0^2 \cdot \frac{1/4}{1/2} + 1^2 \cdot \frac{1/4}{1/2} = 1/8$$

So, the variance is $1/8 - 1/8^2 = 7/64$

(e)

$$E[X^3 + X^2 + 3Y^7|Y = 1] = E[X^3|Y = 1] + E[X^2|Y = 1] + 3 \cdot E[Y^7|Y = 1]$$

Since $0^2 = 0$ and $1^2 = 1$, much of the math from the previous section stays the same resulting in:

$$E[X^3|Y = 1] = {}^1\!/\!_8$$
$$E[X^2|Y = 1] = {}^1\!/\!_8$$

Finally, $\cdot E[Y^7|Y = 1]$ is just the conditional expectation of $Y = 1$ which is ${}^1\!/\!_2$. That means $3 \cdot E[Y^7|Y = 1] = {}^3\!/\!_2$.

Putting it all together, $E[X^3 + X^2 + 3Y^7|Y = 1] = 1.75$.

Problem 2. The projection of vector $\vec{x}$ onto some subspace $\mathbf{V}$ is given by:

$$\text{Proj}_V(\vec{x}) = \boldsymbol{A}(\boldsymbol{A^T}\boldsymbol{A})^{-1}\boldsymbol{A}^T\vec{x} \tag{1}$$

Using numpy, we get $\text{Proj}_V(\vec{P_1}) = [3, 3, 3]$, $\text{Proj}_V(\vec{P_2}) = [1, 2.5, 2.5]$, and $\text{Proj}_V(\vec{P_3}) = [0, .5, .5]$

Problem 3. Because we are looking for the probability that there at most 50 heads, we should use the binomial cumulative distribution function. This particular problem is expressed as

$$Pr(X \leq 50) = \sum_{i=0}^{50} \binom{100}{i} \frac{2}{3}^i \frac{1}{3}^{100-i}$$

We can then use scipy's binomial cumulative distribution function to calculate this. This outputs a .04% chance.